EXHIBIT E

Message

From: Frank Zhang [[redacted]@meta.com]
Sent: 4/16/2024 5:00:51 AM
To: James Beldock [[redacted]@meta.com]; Varun Vontimitta [[redacted]@meta.com]; Joe Spisak [[redacted]@meta.com]; Beto de Paola [[redacted]@meta.com]; Tianhe Li [[redacted]@meta.com]; Patrick Alrassy [[redacted]@meta.com]; Lovish Madaan [[redacted]@meta.com]; Hugo Touvron [[redacted]@meta.com]; Jeet Shah [[redacted]@meta.com]; Rohit Patel [[redacted]@meta.com]; Vedanuj Goswami [[redacted]@meta.com]; Frank Zhang [[redacted]@meta.com]; Rui Hou [[redacted]@meta.com]; Sergey Edunov [[redacted]@meta.com]; Binh Tang [[redacted]@meta.com]; Jeremy Fu [[redacted]@meta.com]; Parth Parekh [[redacted]@meta.com]; Praj Bhargava [[redacted]@meta.com]; Dieuwke Hupkes [[redacted]@meta.com]; Sheng Shen [[redacted]@meta.com]; Dhruv Mahajan [[redacted]@meta.com]; Sharan Narang [[redacted]@meta.com]
Subject: Message summary [{"otherUserFbId":null,"threadFbId":7322926041086533}]
Attachments: 434438572_935385868237358_1784195943901822060_n.png; 436602541_1122676669040122_1660631145520389843_n.jpg; 434806417_710855007928175_8407661496208913503_n.png; 437930594_959133698705606_1909766835560690051_n.png

Patrick Alrassy (4/15/2024 08:01:08 PDT):
>Hi @Frank @Vedanuj , we are reporting in the blog post and model card the final 7B (annealed w/o Math) eval numbers.
>
>dashboard shows we're using `/mnt/wsfuse/outputs/llama3_[redacted]
>
>but this sheet https://docs.google.com/spreadsheets[redacted] looks like there is a `checkpoint_2925000_v11`
>
>
>1- what's the difference ?
>2- what checkpoint we should be using to report evals ?

Vedanuj Goswami (4/15/2024 08:03:16 PDT):
>Frank please correct me if I am wrong, I think
>
>1. annealed_no_math/[redacted] is on v9 annealing mix and checkpoint_2925000_v11 is on v11 annealing mix
>2. we should use the model that was used for later post training stages

Vedanuj Goswami (4/15/2024 08:07:16 PDT):
>@Rui Do you know which 7b annealed model you all ended up using for the post training stages?
>
>`checkpoint_2925000` OR `checkpoint_2925000_v11`?

Vedanuj Goswami (4/15/2024 08:27:15 PDT):
>Confirmed with Sheng .. we used `checkpoint_2925000_v11`. @Patrick we should run the 7B evals on the checkpoint_2925000_v11 checkpoint for pretraining.

Patrick Alrassy (4/15/2024 08:29:07 PDT):
>ok thanks @Vedanuj , I'll rerun

Patrick Alrassy (4/15/2024 08:51:40 PDT):
>sorry team for the pings as we finalize and review.
>@Binh , for `AGI English (3-5 shot, average/acc_char) ` we 're using those subtasks only ... ok ?

Patrick Alrassy (4/15/2024 08:51:40 PDT):

shared: 434438572_935385868237358_1784195943901822060_n.png

Binh Tang (4/15/2024 09:07:18 PDT):
>that's correct

Beto de Paola (4/15/2024 10:47:43 PDT):
>Added the questions about the remaining tests and matching the correct lm harness tests to run in the doc notes as well:
https://docs.google.com/document[redacted]c @Rohit @Jeet

Beto de Paola (4/15/2024 10:48:22 PDT):

>On the checkpoints we converted for testing, Jeet compiled the ones provided in this section:
https://docs.google.com/document/[redacted]

Patrick Alrassy (4/15/2024 11:19:44 PDT):
>a lot of hard work here, thanks @Beto de Paola and @Jeet Shah
>
>qq, where I can check the number of shot configuration and prompt format you passed to the lm_harness , I'd like to investigate some of the differences.
>
>also i highlighted the differences in yellow in the doc (hope thats ok)

Beto de Paola (4/15/2024 11:45:38 PDT):
>The number of shot for each run is in the column for each task and also in the configs that's in the paste for each result.

Beto de Paola (4/15/2024 11:46:13 PDT):
>For the prompt, I'm going to run the remaining tests with the full output for each one, so we can evaluate it. Also will re run DROP and GSM8k so we can evaluate those as well

Rohit Patel (4/15/2024 11:51:55 PDT):
>@Dhruv Mahajan @Rui Hou share RC3 checkpoint asap please. Also, @Dhruv Mahajan did you send the data for human eval over the weekend?

Rui Hou (4/15/2024 12:05:11 PDT):
>Hi @Rohit - I will need to defer this to @Dhruv for the final candidate.

Redacted

Sergey Edunov (4/15/2024 12:55:28 PDT):
>@James Beldock you asked for a fire 🔥 can you help with this one?

Rohit Patel (4/15/2024 12:57:25 PDT):
>@Dhruv Mahajan

Rohit Patel (4/15/2024 12:57:43 PDT):
>Aston confirmed we sent 70B mid for human evals

Rohit Patel (4/15/2024 12:57:51 PDT):
>Is that final we are shipping

Dhruv Mahajan (4/15/2024 13:02:33 PDT):
>https://docs.google.com/spreadsheets/[redacted]

Dhruv Mahajan (4/15/2024 13:02:40 PDT):
>@Rohit Patel two green rows here

Dhruv Mahajan (4/15/2024 13:02:48 PDT):
>#shareadoc

Rohit Patel (4/15/2024 13:03:59 PDT):
>@Sheng Shen do you have headers for those metrics and can you please populate RC3 in L3P sheet?

Tianhe Li (4/15/2024 13:07:12 PDT):
>Hi I'm checking the reported number for the blog post again. qq - Do Math and GSM8k by default use CoT prompt? For example, Mistral, Gemma, and Gemini papers don't explicitly mention using CoT, should we call out them as "Not CoT" in the sheet?

Lovish Madaan (4/15/2024 13:13:07 PDT):

shared: 436602541_1122676669040122_1660631145520389843_n.jpg

Lovish Madaan (4/15/2024 13:13:12 PDT):
>From gemini, they do use CoT

Binh Tang (4/15/2024 13:13:14 PDT):
>In the Gemma paper, they refer to the evaluation setup in the Gemini 1.0 report, which includes CoT for MATH and GSM8K

Frank Zhang (4/15/2024 13:13:29 PDT):
>math and gsm8k come with CoT in the eval prompt. so I assume it's used quite widely every where

Rohit Patel (4/15/2024 13:18:47 PDT):
>yeah so we don't need to call it out for others

Rohit Patel (4/15/2024 13:30:21 PDT):

>@Patrick Alrassy @Lovish Madaan @Binh Tang let's call out everything we need do, especially ARC, the triviaQ wiki split, etc.. where needed we should showcase prompts: https://docs.google.com/document/[redacted]

James Beldock (4/15/2024 13:47:27 PDT):
>Just seeing. Understand. Reading in.

James Beldock (4/15/2024 13:50:40 PDT):
>OK, we've got a small work stream as of this morning working to unblock evals for the forward-looking parts of the blog post (about 405B and multimodal).

Redacted

Sergey Edunov (4/15/2024 13:51:30 PDT):
>no no no, releasing numbers is not an issue

Joe Spisak (4/15/2024 13:51:35 PDT):
>We are NOT making the models available

Sergey Edunov (4/15/2024 13:52:21 PDT):
>we are talking about potentially releasing our model outputs and inputs for specific set of benchmarks, which includes benchmarks themselves

Sheng Shen (4/15/2024 13:53:10 PDT):
>sure, just added them in, let me confirm the metric folder as well and will put there once confirmed.

Sergey Edunov (4/15/2024 13:53:23 PDT):
>I think... this is most likely okay, but e.g. if benchmarks present copyrighted material, then I suppose we don't want to release them (even if we used them/report numbers on them/and they are available online)

James Beldock (4/15/2024 13:53:42 PDT):
>Ah hah, OK.

James Beldock (4/15/2024 13:53:54 PDT):
>Thank you. Who are the 2-3 people on this team who can represent the requirements?

Redacted

Sergey Edunov (4/15/2024 13:55:46 PDT):
>@Patrick, @Rohit? not sure how to prioritize it, since we haven't yet decided that we will need it. Rohit you're running the HF harness, what's the likelihood that we will need it by Thursday?

Redacted

Patrick Alrassy (4/15/2024 14:10:13 PDT):
>thanks @James

Rohit Patel (4/15/2024 14:11:42 PDT):
>I am hoping this is low and that we don't have to do this. We should understand what it would take to do it either way. I think for future releases this is a pretty good idea TBH

Rohit Patel (4/15/2024 14:13:20 PDT):
>@Beto de Paola when do we think we will have all the numbers and for the areas that we do have differences who is taking point on looking at the responses and figuring out what is causing the delta

Rohit Patel (4/15/2024 14:14:02 PDT):
>None of this should take away from #1 priority which is to complete that sheet with correct (and matching) configs for all numbers and I assume @Patrick Alrassy you are on point for that

Patrick Alrassy (4/15/2024 14:14:24 PDT):
>yes

Beto de Paola (4/15/2024 14:18:40 PDT):
>Estimating by EOD today, PT time. Had some issues with DROP and lost both runs on 70b and 8b (rerun).

Beto de Paola (4/15/2024 14:21:34 PDT):
>Shared with @Patrick Alrassy the re run of GSM8K on 8b. It was even worse than the first run.

Sergey Edunov (4/15/2024 14:29:03 PDT):
>what are we getting? sorry, is there a place where we track the numbers?

Patrick Alrassy (4/15/2024 14:29:53 PDT):
>https://docs.google.com/document/d/[redacted]

Patrick Alrassy (4/15/2024 14:31:14 PDT):
>this bookmark @Sergey Edunov
>
>https://docs.google.com/document/d/[redacted]

Lovish Madaan (4/15/2024 14:46:03 PDT):
>can you increase the max generation length? It seems the model is not getting the chance to generate a final answer and is cut off midway through generation

Rohit Patel (4/15/2024 14:53:21 PDT):
>Can we please add a column for this run in the L3P Data sheet so we can look at it all easily

Patrick Alrassy (4/15/2024 14:56:44 PDT):
>yes I agree. @Beto de, looks like from https://www.internalfb.com/intern/[redacted] the output is truncated leading the parser to parse the wrong index. We should increase the model maximum generation length parameter.
>@Beto de do you know how to do that ? or need help from others here ?

Beto de Paola (4/15/2024 14:57:16 PDT):
>I'll check quickly

Beto de Paola (4/15/2024 15:18:10 PDT):
>We can re run this with increased gen size, what value should we set?

Beto de Paola (4/15/2024 15:18:28 PDT):
>Also, which other tests might be affected by this? arc_challenge?

Patrick Alrassy (4/15/2024 15:24:32 PDT):
>for generation tasks, we should set `max_gen_len=512`
>for choice tasks (multiple choice tasks) , we should set `max_gen_len=10`
>
>@Frank Zhang @Binh Tang plz kmh here

Patrick Alrassy (4/15/2024 15:25:34 PDT):
>>Also, which other tests might be affected by this?
>
>potentially any task that is not multiple choice.
>
>also @Frank @Binh to confirm plz #silent

Beto de Paola (4/15/2024 15:43:40 PDT):
>Double checking this, these are the latest checkpoints to test on?

Jeet Shah (4/15/2024 15:43:53 PDT):
>@Dhruv Mahajan these are RC2 candidates

Jeet Shah (4/15/2024 15:43:58 PDT):
>Do we want to move to RC3 candidates

Patrick Alrassy (4/15/2024 16:19:43 PDT):
>Following up on why we see huge differences between our ARC-C (pretrain & posttrain) and lm_harness ones.
>
>Tl,dr: we do generation , they do choice. We present all choices in the prompt/few shot, they present 1 shot. Those are things we already in the past observed that we regress on if we present the task and prompt like they did (as far as I remember cc: @Binh Tang right ?)
>
>For pretrain: we & lm_harness run ARC-C as a NLL ChoiceTask , but
>
>Our prompt : we enumerate all the choices in the prompt each time we formulate the 4 NLL prompts. And what we vary is the answer letter only Answer: A, Answer: B … .
>https://github.com/[redacted]
>
>Harness prompt: they dont submit all the choices in their few shot/prompt. But rather only the correct answer. I.e they dont run it like they run MMLU. (see here https://www.internalfb.com/[redacted])
>
>So I think when we run our model against harness we regressed based on that.
>
>
>
>

>
[redacted]
>
>
[redacted]
> {
[redacted]
> },
>```
>
>
>
>
>Our Finetuned prompt: it's 1 prompt, we present the 4 possible answer. And we let the model Generate the best answer.
>https://github.com/fairinternal[redacted]
>
>
>cc: @Rohit @Sergey @Binh @Beto de

Patrick Alrassy (4/15/2024 16:19:43 PDT):

shared: 434806417_710855007928175_8407661496208913503_n.png

Patrick Alrassy (4/15/2024 16:21:27 PDT):
>Following up on why we see huge differences between our ARC-C (pretrain & posttrain) and lm_harness ones.
>
>Tl,dr: we do generation , they do choice. We present all choices in the prompt/few shot, they present the correct answer only. Those are things we already in the past observed that we regress on if we present the task and prompt like they did (as far as I remember cc: @Binh Tang right ?)
>
>For pretrain: we & lm_harness run ARC-C as a NLL ChoiceTask , but
>
>Our prompt : we enumerate all the choices in the prompt each time we formulate the 4 NLL prompts. And what we vary is the answer letter only Answer: A, Answer: B … .
>https://github.com/fairinternal/evals/[redacted]
>
>Harness prompt: they dont submit all the choices in their few shot/prompt. But rather only the correct answer. I.e they dont run it like they run MMLU. (see here
https://www.internalfb.com/[redacted]
>
>So I think when we run our model against harness we regressed based on that.
>
>
>
>
>
[redacted]
>
>
[redacted]
>E.g. [redacted]
>
> [redacted]

```
> [redacted]
> [redacted]
> [redacted]
>
>
>
>
>
>
>    },
>
>
>
>
>Our Finetuned  prompt:  it's 1 prompt, we present the 4 possible answer. And we let the model Generate the best answer.
>https://github.com/fairinternal/[redacted]
>
>cc: @Rohit Patel @Sergey Edunov @Binh Tang @Beto de Paola
```

Patrick Alrassy (4/15/2024 16:21:28 PDT):

shared: 437930594_959133698705606_1909766835560690051_n.png

Rohit Patel (4/15/2024 16:21:48 PDT):
>Does this make a material difference. Given the timeline and if we already have RC2 we can go with that

Frank Zhang (4/15/2024 16:24:29 PDT):
>The observation here is consistent with Binh's analysis here:
https://fb.workplace.com/[redacted]

Sheng Shen (4/15/2024 16:25:00 PDT):
>hey Jeet, here are RC3 candidates
>
>70B: https://docs.google.com/spreadsheets/[redacted]
>
>7B: https://docs.google.com/spreadsheets/d/[redacted]
>
>#sharedoc edit

Beto de Paola (4/15/2024 16:39:45 PDT):
>Re running drop for 8b, and GSM8k for both 8b and 70b. Should be done in about 3 hs total. Updated the doc with the latest.

Rohit Patel (4/15/2024 21:53:36 PDT):
>Hey all, can we find time first thing tomorrow to discuss these results and mitigations? We are eliminating NQ and Hellaswag from the model card so @Beto de Paola you can kill those jobs. Please also ignore TQA becuse the split is different (I have half the mind to kill that too). Do we have an investigation on GSM8K delta, and the weird numbers I am seeing on DROP?

Beto de Paola (4/15/2024 21:59:56 PDT):
>@silent On DROP: I'm uploading the last results I got shortly and will update the doc. It seems the model is answering correctly, but appending additional information into the answer.

Beto de Paola (4/15/2024 22:00:46 PDT):
>@silent On GSM8K, had some hiccups running it and the last test just finished on 8B, but it's not better than before. Will also update the doc.

Rohit Patel (4/15/2024 22:00:51 PDT):
>I've put a meeting on the calendar morning 9AM PT. @Binh Tang @Lovish Madaan does that work for you as I'm seeing some conflicts on your calendar